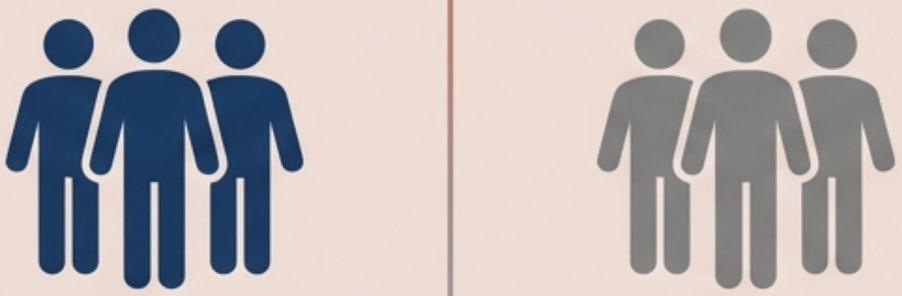
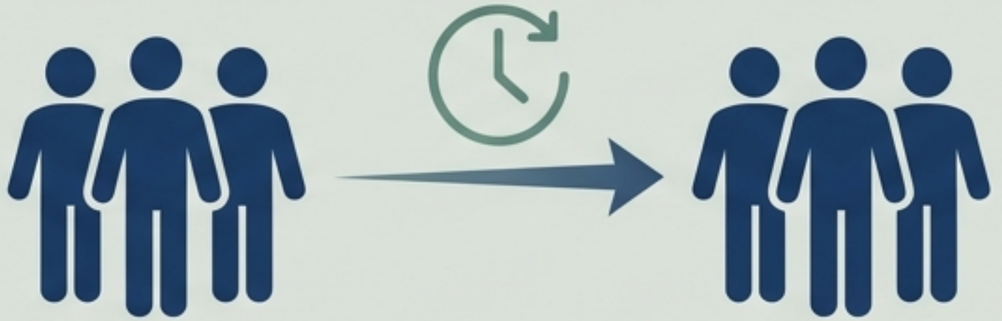
Comparing Two Means in Movement Science

Independent



Paired



The Core Question

Is there a meaningful difference between these two groups, or could it simply be sampling variability?

The T-Test Framework

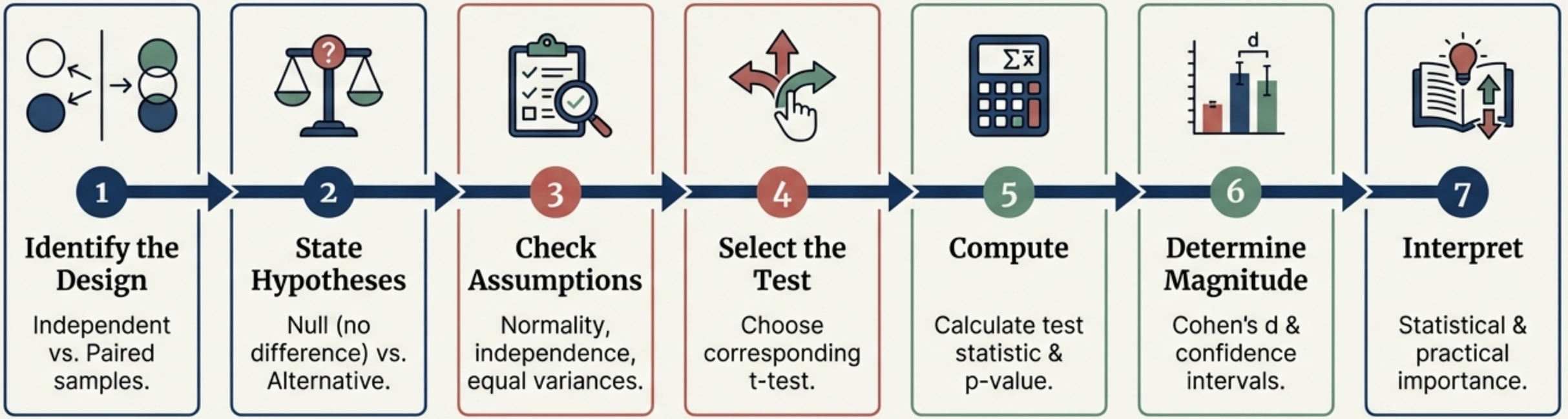
A principled statistical method accounting for sample size and data variability to yield p-values. It quantifies evidence against the null hypothesis (no difference).

The Goal

Move beyond binary significant/non-significant thinking by integrating hypothesis testing with estimation, effect sizes, and practical significance.

NotebookLM

The 7-Step Workflow for Mean Comparisons



Software Note: While conceptual equations are vital, modern researchers rely on analytical software (like SPSS) to compute these steps, avoiding tedious by-hand calculation errors.

NotebookLM

Evaluating Separate Groups: The Independent T-Test

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{SE_{diff}}$$

Where:

$\bar{x}_1$, $\bar{x}_2$ = sample means for Groups 1 and 2

SE_{diff} = standard error of the difference between means

Equal variances assumed (pooled variance)

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$SE_{diff} = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$df = n_1 + n_2 - 2$$

Equal variances not assumed (Welch's t-test)

$$SE_{diff} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$df \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 (n_1\text{-}1) + \left(\frac{s_2^2}{n_2}\right)^2}$$

Application

Comparing two separate, unrelated groups (e.g., VO2max in 25 athletes vs. 25 non-athletes). Each participant contributes one score.

Hypotheses

H_0: $\mu_1 = \mu_2$ vs. H_1: $\mu_1 \neq \mu_2$ (Two-tailed preferred).

Welch's T-Test Default

The Welch's calculation is the modern, robust default in software as it does not require the equal variance assumption and performs well even when variances are equal.

Worked Example

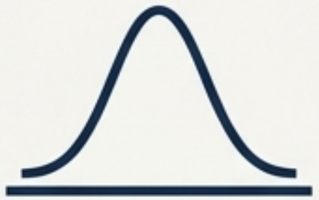
Young adults (n=20) vs. Older adults (n=18) reaction time. Software shows $p < .001$, concluding young adults had significantly faster reaction times.

NotebookLM

Validating Independent T-Test Assumptions

Independence of Observations

Scores in one group absolutely cannot influence scores in the other group.



Normality

Data in each group should be roughly normal. Check visually (Histograms, Q-Q plots) or via formal tests (Shapiro-Wilk for most sizes). Robust to moderate non-normality if n > 30 per group.

Homogeneity of Variance

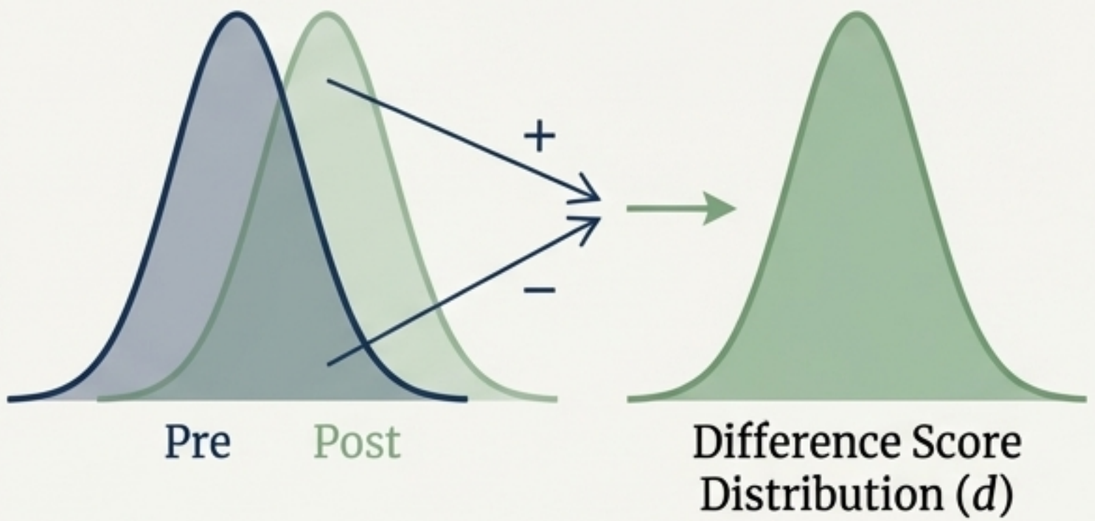
Check via Levene's Test (H0: variances are equal). If the ratio of larger to smaller ... If the ratio variance is < 2:1, the assumption is reasonable.

Common Mistake: Assuming normality for small samples

With n < 15 per group, moderate departures from normality ruin validity. Check visually and consider nonparametric alternatives (Mann-Whitney U) if violated.

NotebookLM

Analyzing Related Measurements: The Paired T-Test



Application

Repeated measures (pre/post-test) or matched pairs (twins, left vs. right limb). Each participant or pair contributes one difference score.

The Power Advantage

Removes between-subject variability, leaving only within-subject variability. This controls for individual differences and drastically increases statistical power.

Hypotheses: H_0: $\mu_d = 0$ vs. H_1: $\mu_d \neq 0$.

Mathematically equivalent to a one-sample t-test on the differences.

Paired T-Test

$$t = \frac{\bar{d} - 0}{SE_d} = \frac{\bar{d}}{s_d/\sqrt{n}}$$

where $d = x_{i,\text{after}} - x_{i,\text{before}}$

Worked Example

Grip strength in 12 climbers measured before and after a 6-week program. Software automatically assesses d = post - pre.

NotebookLM

Validating Paired T-Test Assumptions

WRONG: Testing raw pre-test and post-test scores separately for normality.

RIGHT: Testing the calculated difference scores (d = post - pre) for normality.

Pairs are Independent

One matched pair or individual's scores must not influence another's.

Differences are Normally Distributed

Check the normality of the calculated difference scores, absolutely not the raw scores.

No Order Effects

For repeated measures, experimental design elements like counterbalancing or randomization must prevent systematic order effects (e.g., fatigue or learning).

Core Principle: The paired t-test strictly analyzes differences; therefore, only the distribution of those differences matters for statistical validity.

NotebookLM

Decision Matrix: Independent vs. Paired Designs

Characteristic	Independent t-test	Paired t-test
Design	Two separate groups	Same participants measured twice
Assumptions	Independence, normality, equal variances	Pairs independent, differences normally distributed
Degrees of Freedom	$n1 + n2 - 2$ (or Welch's df)	$n - 1$ (where n is number of pairs)
Example	Trained vs. Untrained	Pre-test vs. Post-test
Null Hypothesis	$\mu1 = \mu2$	$\mu d = 0$

⚠ Critical Penalty

Using an independent t-test when data are actually paired results in a severe loss of power by failing to control for individual differences. Always match the test to the design.

NotebookLM

Quantifying Magnitude: Effect Sizes and Confidence Intervals

Cohen's *d* for Independent Samples

$$s_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{\text{pooled}}}$$

Cohen's *d* for Paired Samples

$$d = \frac{\vec{d}}{s_d}$$

Where $\vec{d}$ is the mean difference and sd is the SD of differences.

Beyond P-Values

Statistical significance merely tells us an effect is detectable; effect sizes tell us how large the effect is in standardized units.

Context Matters

Benchmarks are guidelines. A 'small' effect (d = 0.2) can be highly practical in injury prevention, while d = 1.42 (young vs. older adult reaction times) is substantial.

Confidence Intervals

Effect sizes have uncertainty. Wide CIs indicate low precision.
Example: Resistance training improves strength with d = 0.78, 95% CI [0.65, 0.91].

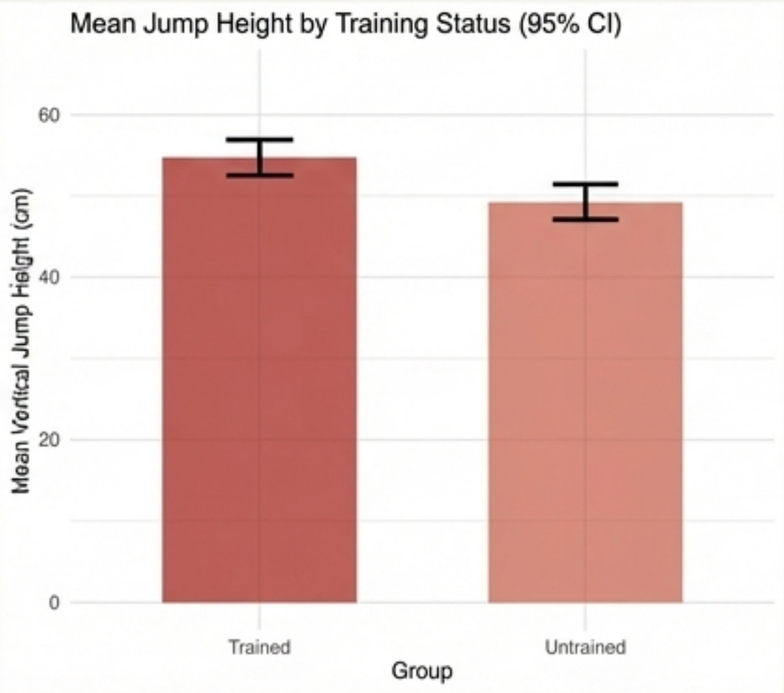
NotebookLM

Best Practices for Visualizing Group Comparisons

Box Plots (Independent)

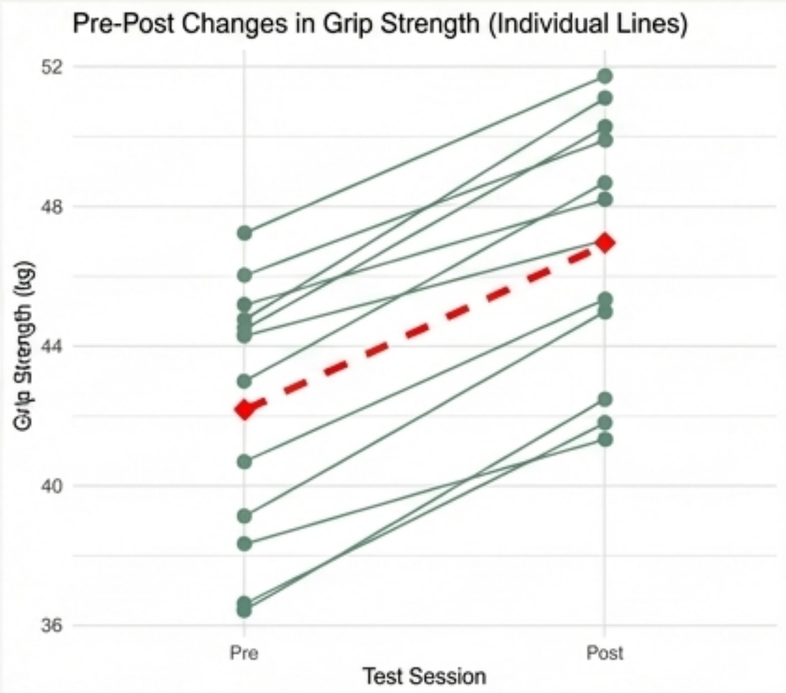
Reveal the distribution, medians, IQR, and outliers within each separate group.

Error Bar Plots



Provide a visual comparison of means and precision. Non-overlapping 95% CIs typically indicate roughly $p < .01$, but never rely on visuals alone for significance.

Trajectory Plots (Paired)



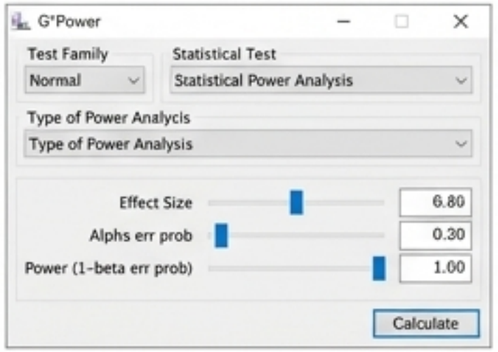
Crucial for repeated measures. Displays how every individual participant changed alongside the mean change (dashed line), highlighting both overall trend and individual variability.

NotebookLM

Ensuring Detection: Sample Size and Statistical Power

Statistical Power: The probability of detecting a true effect when it exists. Underpowered studies waste resources and produce false negatives. Target power is typically ≥ 0.80. Four Factors: Sample size (n ↑), Effect size (d ↑), Significance level (α ↑), and Design (Paired > Independent).

Best Practice: A Priori Planning



Use software like G*Power before data collection to determine the exact required sample size based on expected effect size and desired power.

Warning: The Post-Hoc Fallacy

Computing power after data collection using the **observed effect size** is **mathematically circular**. A **non-significant p-value guarantees low observed power. Report effect sizes with CIs** instead.

NotebookLM

Navigating Assumption Violations and Alternatives

Violation: Severe Non-Normality
Severe non-normality occurs when the data distribution significantly deviates from normal, potentially skewing parametric test results.

→

Solution: Data Transformations
Data Transformations (Log, square root, rank)
OR
Nonparametric alternatives (Mann-Whitney U for independent, Wilcoxon signed-rank for paired).

Violation: Unequal Variances
Unequal variances (heteroscedasticity) violate the assumption that groups have similar variability, affecting the validity of standard t-tests.

→

Solution: Welch's t-test. This is the preferred default for independent samples as it entirely bypasses the homogeneity assumption.

Violation: Very Small Samples (n < 15)
Very small samples (n < 15) often lack the power to reliably detect effects and are highly sensitive to outliers.

→

Solution:
Normality is highly suspect. Consider bootstrapping or permutation tests for robust inference.

NotebookLM

Comparing Categorical Rates: The Two-Proportion Z-Test

Z-Test Statistic (Unpooled Standard Error)

$$z = \frac{p_1 - p_2}{SE_{diff}}$$

$$SE_{diff} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Pooled Proportion ($\hat{p}$) and its Standard Error

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$SE_{diff(\text{pooled})} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Application

Used when comparing proportions, such as injury rates or success rates, rather than continuous means. Follows a standard normal (z) distribution for large samples.

Unpooled vs. Pooled

Unpooled standard error is used for computing confidence intervals. Pooled proportion (p̂) is theoretically preferred for formal significance testing under H0: p1 = p2.

Injury Rate Example

18% injuries in minimalist shoes vs. 12% in traditional shoes. Result yields z = 1.20, p = .23, indicating no statistically significant difference in injury rates.

NotebookLM

Reporting Results in Strict APA Style

Independent T-Test Template

```
[Group 1] (M = x, SD = x) differed significantly from [Group 2] (M = x, SD = x), t(df) = x, p = x, d = x, 95% CI [lower, upper].
```

Example: Trained athletes (M = 55.2, SD = 4.1) demonstrated significantly higher vertical jump performance than untrained controls (M = 47.7, SD = 5.8), t(58) = 4.52, p < .001, d = 1.12, 95% CI [4.2, 10.8] cm.

Paired T-Test Template

```
[Condition 2] was significantly higher than [Condition 1]... t(df) = x, p = x, mean difference = x, 95% CI [lower, upper], d = x.
```

Example: Post-training grip strength (M = 51.5, SD = 2.8) was significantly greater than pre-training grip strength (M = 47.2, SD = 3.1), t(11) = 16.91, p < .001, mean difference = 4.3 kg, 95% CI [3.77, 4.90], d = 4.87.

NotebookLM

Common Pitfalls and Best Practices

Pitfall

P-hacking and selective reporting (inflates Type I error).

Solution

Preregister analyses, report all tests, correct for multiple comparisons.

Pitfall

Confusing significance with importance (trivial effects trigger $p < .05$ in large samples).

Solution

Always report and interpret effect sizes (d) and confidence intervals.

Pitfall

Ignoring assumptions (yields incorrect p-values).

Solution

Check visual/formal assumptions, default to Welch's t-test, or use nonparametrics.

Pitfall

Inappropriate test choice (using independent test for paired data).

Solution

Strictly match the statistical test to the research design to preserve power.

NotebookLM

Core Takeaways for Movement Science Researchers

- ✓ Design Dictates Method: Independent tests compare separate groups; paired tests compare related measurements (offering greater power).
- ✓ Robust Defaults: Always check normality and independence. Use Welch's t-test to bypass homogeneity of variance issues.
- ✓ Plan for Detection: Adequately powered studies require a priori planning.

The Nuanced Approach

"Responsible research requires integrating hypothesis testing with estimation. Do not merely ask "Is there a significant difference?" Ask "How large is the difference, how precisely is it estimated, and does it practically matter?"

NotebookLM